EFFICIENT SCREENING OF VARIABLE SUBSETS IN MULTIVARIATE STATISTICAL MODELS

A. Pedro Duarte Silva

Faculdade de Ciências Económicas e Empresariais,

Universidade Católica Portuguesa - Centro Regional do Porto

Rua Diogo Botelho, 1327

4150 PORTO PORTUGAL

Email: psilva@porto.ucp.pt

In applied statistical studies, it is common to collect data on a large pool of candidate variables from which a small subset will be selected for further analysis. The practice of variable selection often combines the use of substantive knowledge with subjective judgment and data-based selection procedures. The most popular of such procedures are stepwise methods. However, stepwise selection methods have two fundamental shortcomings:

a) Most theoretical results from classical statistics require the assumption that the set of variables to be analyzed was chosen independently of the data. Therefore, when the variables are selected based on the data, the results from classical distribution theory almost never hold.
b) Stepwise selection methods look at one variable at a time and tend to ignore the impact of combining particular sets of variables together. Thus, as each variable "importance" is often influenced by the set variables currently under analysis, stepwise methods may fail to identify the most adequate variable subsets.

The problems created by a) and b) are now widely recognized and have been discussed by several authors. Miller (1984, 1990) and Derksen and Keselman (1992) give good reviews of the relevant literature in the context of Regression Analysis. In the context of Discriminant Analysis the problems created by a) are discussed, among others, by Murray (1977), McKay and Campbel (1982a), Snapinn and Knoke (1989) and Turlot (1990), and the problems created by b) are discussed by Hand (1981), McKay and Campbel (1982a, 1982b) and Huberty and Wisenbaker (1992).
The problem referred in b) can be overcome by procedures that compare all possible variable subsets according to appropriate criteria. However, this approach usually requires the evaluation of a large number of alternative subsets, and may not be feasible. For regression models, several efficient algorithms were developed in order to surpass this problem. For instance, for the linear regression model with p candidate variables, Beale, Kendall and Mann (1967) and Hocking and Leslie (1967) proposed branch and bound algorithms that identify "the best" (in the sense of R^2) variable subsets, evaluating only a small fraction of the 2^p-1 different subsets. Furnival (1971) has shown how the residual sum of squares for all possible regressions can be computed with an effort of about six floating point operations per regression. Furnival and Wilson (1974) combined Furnival algorithm with a branch and bound procedure, leading to the widely used "leaps and bounds" algorithm for variable

selection. Lawless and Singhal (1978) adapted Furnival and Wilson algorithm to non-linear regression and Kuk (1984) applied the former adaptation to proportional hazard models in survival analysis.

This article will focus on efficient algorithms for all-subsets comparisons in multivariate analysis. In particular, it will be shown how Furnival and Wilson "leaps and bounds" can be adapted to compare, according to appropriate criteria, variable subsets in Discriminant Analysis, MANOVA, MANCOVA, and Canonical Correlation Analysis. To the best of our knowledge, algorithms to surpass the difficulties created by b) have not received the same amount of attention in multivariate models as in regression. Some exceptions include the following research. McCabe (1975) adapted Furnival's algorithm to the comparison (according to Wilk's Λ) of variable subsets in Discriminant Analysis. McHenry (1978) proposed a compromise between stepwise and all-subsets procedures in multivariate linear models. Seber (1984, pp 507-510) discusses extensions of McCabe approach to variable comparisons concerning linear hypothesis in multivariate models, and briefly mentions that branch and bound algorithms can also be employed. Duarte Silva (forthcoming) adapted Furnival and Wilson "leaps and bounds" to the minimization of parametric estimates of the error rate in two-group Discriminant Analysis.

No attempt will be made here to deal with the difficulties referred in a). Although any data-based variable selection procedure usually leads to violations of the assumptions underlying classical inference methods, that should be no reason for ignoring the data in the variable selection process. Anyway, for the purpose of statistical inference it is not recommended that the effects of variable selection should be ignored. When inference is required and the variables are not chosen a-priori, specialized procedures should be employed. Some possibilities in that regard, include the use of cross-validation techniques, like bootstrap methodologies that explicitly take into account the selection process (ex: Snapinn and Knoke 1989).

All the procedures discussed in this article are being programmed in the C^{++} language. A public-domain software implementation for Personal Computers can be ordered directly from the author and will soon be available on the internet at the address http://www.porto.ucp.pt/~psilva.

OUTLINE OF "LEAPS AND BOUNDS" ALGORITHMS

Most modern algorithms for all-subsets comparisons in statistical models are based on adaptations of Furnival (1971) and Furnival and Wilson (1974) algorithms for variable selection in linear regression. Within a general framework, these algorithms may be outlined as follows. Assume that there are p candidate variables, $\mathbf{X_1, X_2, ..., X_p}$ to enter a given statistical model. Denote the different subsets of $\{\mathbf{X_1, X_2, ..., X_p}\}$ by $\mathbf{S_1, S_2, ..., S_2P_{-1}}$, where $S_1 = X$ represents the full set comprising all p candidates. We are interested in the comparison of $\mathbf{S_1, S_2, ..., S_2P_{-1}}$, according to appropriate criteria of "model quality". Suppose that it is possible to define one such criterion, $C(S_a)$, that can be expressed as a function of a quadratic form, $Q(S_a)$, with general expression $Q(S_a) = \mathbf{v_{S_a}}' \mathbf{M^{-1}_{S_aS_a}} \mathbf{v_{S_a}}$.

The following notation is adopted through this text. Matrices are denoted by bold upper case and vectors by bold lower case. Individual elements of vectors and matrices are denoted by lower case subscripts. The portions of vectors and matrices associated with the variables included in a given subset, S_a, are denoted by the subscripts S_a (vectors) and S_aS_a (matrices).

The complement of S_a is denoted by $\bar{S}_a$ and the matrices comprised by the rows (columns) associate with variables included in S_a and columns (rows) associated with variables excluded from S_a are denoted by the subscripts $S_a \bar{S}_a$ ($\bar{S}_a S_a$). The i-th row (j-th column) of a matrix associated with the variables included in S_a is denoted by the subscripts i,S_a (S_a,j).
Assume that there are a column vector $\mathbf{v}_{S_1}$ and a symmetric matrix $\mathbf{M}_{S_1S_1}$ satisfying the following conditions:

(A) The (i,j) element of $\mathbf{M}_{S_1S_1}$, $\mathbf{M}_{ij}$, is a function of the values of X_i and X_j. $\mathbf{M}_{ij}$ is not influenced by any variable other than X_i and X_j.

(B) The i-th element of $\mathbf{v}_{S_1}$, $\mathbf{v}_i$, is a function of the values of X_i. $\mathbf{v}_i$ is not influenced by any variable other than X_i.

Then, Furnival algorithm is essentially a method for evaluating all forms $Q(S_a)$ with minimal computational effort. In the linear regression model $\mathbf{y = X\beta + u}$,
a natural comparison criterion is the Residual Sum of Squares, $RSS_{S_a} = \mathbf{y'\ y} - \mathbf{y'X}_{S_a}(\mathbf{X}_{S_a}'\mathbf{X}_{S_a})^{-1}\mathbf{X}_{S_a}'\mathbf{y}$ which can be evaluated by choosing $\mathbf{M}_{S_1S_1}$ and $\mathbf{v}_{S_1}$ respectively as the matrix of Sums of Squares and Cross Products (SSCP) among the regressors and vector of sums of crossproducts between the regressors and the dependent variable, i.e., $\mathbf{M}^{(1)}_{S_1S_1} = \mathbf{X'X}$ and $\mathbf{v}^{(1)}_{S_1} = \mathbf{X'\ y}$. In that case $C^{(1)}(S_a) = RSS_{Sa} = \mathbf{y'\ y} - Q^{(1)}(S_a)$. However, in most implementations of the algorithm, for reasons of numerical stability, these sums are replaced by correlations, defining $\mathbf{M}_{S_1S_1}$ and $\mathbf{v}_{S_1}$ instead as $\mathbf{M}^{(2)}_{S_1S_1} = \mathbf{R_{XX}}$, $\mathbf{v}^{(2)}_{S_1} = \mathbf{r_{Xy}}$. This choice leads to the equivalent criterion $C^{(2)}(S_a) = R^2_{S_a} = \mathbf{r}_{X_{S_a}y}'\,\mathbf{R}^{-1}_{X_{S_a}X_{S_a}}\mathbf{r}_{X_{S_a}y} = Q^{(2)}(S_a)$.
The fundamental mechanics of the algorithm are as follows. Initially, create the source matrix $\mathbf{MV}(\emptyset)$, associated with the empty subset, where $\mathbf{M}_{S_1S_1}$ and $\mathbf{v}_{S_1}$ should satisfy (A) and (B).

$$\mathbf{MV}(\emptyset) = \begin{bmatrix} \mathbf{M}_{S_1S_1} & \mathbf{v}_{S_1} \\ \mathbf{v}'_{S_1} & 0 \end{bmatrix} \qquad (1)$$

Next, start "bringing variables into the model" by performing specialized Gauss-Jordan elimination operations known as "symmetric sweeps" (Beale, Kendall and Mann 1967). After "sweeping" all the elements of S_a, $\mathbf{MV}(\emptyset)$ is converted into matrix $\mathbf{MV}(S_a)$, where, without lack of generality, it is assumed that the rows and columns associated with S_a are placed before those associated with $\bar{S}_a$.

$$\mathbf{MV}(S_a) = \begin{bmatrix} -\mathbf{M}^{-1}_{S_aS_a} & -\mathbf{M}^{-1}_{S_aS_a}\mathbf{M}_{S_a\bar{S}_a} & -\mathbf{M}^{-1}_{S_aS_a}\mathbf{v}_{S_a} \\ -\mathbf{M}_{\bar{S}_aS_a}\mathbf{M}^{-1}_{S_aS_a} & \mathbf{M}_{\bar{S}_a\bar{S}_a} - \mathbf{M}_{\bar{S}_aS_a}\mathbf{M}^{-1}_{S_aS_a}\mathbf{M}_{S_a\bar{S}_a} & \mathbf{v}_{\bar{S}_a} - \mathbf{M}_{\bar{S}_aS_a}\mathbf{M}^{-1}_{S_aS_a}\mathbf{v}_{S_a} \\ -\mathbf{v}_{S_a}'\mathbf{M}^{-1}_{S_aS_a} & \mathbf{v}_{\bar{S}_a}' - \mathbf{v}_{S_a}'\mathbf{M}^{-1}_{S_aS_a}\mathbf{M}_{S_a\bar{S}_a} & -\mathbf{v}_{S_a}'\mathbf{M}^{-1}_{S_aS_a}\mathbf{v}_{S_a} \end{bmatrix} \qquad (2)$$

As $Q(S_a) = -1*\mathbf{MV}(S_a)_{p+1,p+1}$ the comparison criterion $C(S_a)$ can be updated after each sweep. A symmetric sweep on variable $X_k \in \bar{S}_a$, uses equations (3) through (5) to convert $\mathbf{MV}(S_a)$ into $\mathbf{MV}(S_b)$ with $S_b = S_a \cup \{X_k\}$.

$$\mathbf{MV}(S_b)_{kk} = -1 / \mathbf{MV}(S_a)_{kk} \qquad (3)$$

$$\mathbf{MV}(S_b)_{kj} = \mathbf{MV}(S_b)_{jk} = \mathbf{MV}(S_b)_{kk} * \mathbf{MV}(S_a)_{kj} \qquad (j \neq k) \qquad (4)$$

$$\mathbf{MV}(S_b)_{ij} = \mathbf{MV}(S_b)_{ji} = \mathbf{MV}(S_a)_{ij} + \mathbf{MV}(S_b)_{kj} * \mathbf{MV}(S_a)_{ik} \qquad (i \neq k, j \neq k) \qquad (5)$$

Furnival minimizes computational effort by taking advantage of the following facts. As the symmetry of the matrices **MV**(.) is always preserved, only elements on or above their diagonals need to be stored and updated. Furthermore, $\mathbf{MV}(S_b)_{p+1,p+1} = \mathbf{MV}(S_a)_{p+1,p+1} - \mathbf{MV}(S_a)^2_{k,p+1} / \mathbf{MV}(S_a)_{kk}$ and each update of $Q(S_a)$ requires only two floating point operations (multiplications and divisions). When $\mathbf{MV}(S_b)$ is not used to evaluate other subsets, all its remaining elements can be ignored. When $\mathbf{MV}(S_b)$ is used to evaluate other subsets involving at most *t* additional variables, then the elements associated with these variables also need to be updated. In that case, a sweep requires (t+1)*(t+4)/2 floating point operations. By ordering the evaluation of subsets in a appropriate manner it is possible to evaluate all forms $Q(S_a)$ is such a way that: (i) The value of each $Q(S_b)$, can be derived from a sub-matrix of $\mathbf{MV}(S_a)$, where S_a has exactly one less variable than S_b. (ii) Only *p* (portions of) matrices **MV**(.) need to be kept simultaneously in memory. (iii) The number of different sub-matrices that are used in the evaluation of subsets with (at most) *t* additional variables (t=0,1,...,p-1) equals 2^{p-t-1}.

A remarkable consequence of (iii), is that a full (p+1)*(p+1) matrix sweep never needs to be performed, a p*p sweep needs to be performed only once, a (p-1)*(p-1) sweep twice, and (following the same pattern) approximately half (more precisely 2^{p-1}) of the **MV**(.) matrices are not used in the evaluation of additional subsets. Furnival and Wilson (1974) show that the total number of floating point operations used in Furnival algorithm equals $6(2^p) - p(p+7)/2 - 6$, and that it is not possible to compute all $Q(S_a)$ with fewer operations.

McCabe (1975) has shown that Furnival algorithm can be adapted to evaluate criteria based on determinants. In particular, if $\mathbf{M}_{S_1S_1}$ is a matrix satisfying condition (A), then all determinants $|\mathbf{M}_{S_aS_a}|$ can be computed using the following algorithm. Create the initial matrix $\mathbf{MD}(\emptyset) = \mathbf{M}_{S_1S_1}$. Initialize the auxiliary variable $D(\emptyset) = 1$. Proceed as in the original algorithm, and each time $\mathbf{MD}(S_a)$ is updated to $\mathbf{MD}(S_b)$ by sweeping on X_k, also compute $D(S_b) = D(S_a) * \mathbf{MD}(S_a)_{kk}$.

As $|\mathbf{M}_{S_bS_b}| = |\mathbf{M}_{S_aS_a}| * (\mathbf{M_{kk}} - \mathbf{M}_{k,S_a}\ \mathbf{M}^{-1}_{S_aS_a}\ \mathbf{M}_{S_a,k})$ it follows that this procedure will ensure that $D(S_a) = |\mathbf{M}_{S_aS_a}|$ for all non-empty subsets S_a. The adaptation described above was used by McCabe to compare variable subsets in Discriminant Analysis and, as it will be discussed in the following sections, has wide applicability in several multivariate models.

Although when it is desired to evaluate all alternative subsets, it is not possible to improve upon the efficiency of Furnival's algorithm, in practice often it is only required the identification of "good" subsets for further inspection. In order to select the best variable subsets according to a given criterion, it is possible to employ search procedures that are able to recognize that many subsets will never be selected before evaluating them. That is the philosophy behind "branch and bound" algorithms for variable selection. Maybe the best known of these algorithms is the "leaps and bounds" algorithm of Furnival and Wilson for

linear regression. Furnival and Wilson algorithm is based on the following properties: (i) The Residual Sum of Squares of a given model can never decrease with the removal of variables, i.e., $S_a \subset S_b \Rightarrow$ $RSS_{S_a} \leq RSS_{S_b}$. ii) Symmetric sweeping is reversible. In effect, if the elements of **MV**(.) are multiplied by minus one, then the symmetric sweep operator updates the resulting matrices when variables are removed. Creating initially the matrices **MV**(Ø) and **MV**(S_1) associated respectively with the empty and full subsets, Furnival and Wilson build a search tree that on the left side moves from **MV**(Ø) adding variables to previous subsets and on the right side moves from **MV**(S_1) removing variables. Finding on the left side of the search tree "good" subsets early on, it is often possible to prune large branches of the search tree that given (i) can never include subsets "deserving" to be selected. In their original article, Furnival and Wilson describe some "smart" strategies to conduct the search, that attempt to maximize pruning, while ensuring that in the worst case (no pruning) the number of floating point operations approaches six per subset.

The basic ideas behind Furnival and Wilson algorithm are not restricted to linear regression models. In fact, they can be applied in any model where it is possible to define criteria that never improve with the removal of variables and operators to reevaluate these criteria upon the addition or removal of single variables. The computational efficiency of such procedures is bound to be dependent on the effort required by these operators. For criteria derived from quadratic forms based on matrices and vectors satisfying (A) and (B), Furnival strategy can be used and the reevaluation of the appropriate criteria is computationally "cheap". The same applies to criteria derived from determinants based on matrices satisfying (A), because the operator used in McCabe's adaptation of Furnival algorithm is also reversible, i.e., if $D(S_b) = |\mathbf{M}_{S_bS_b}|$, then $D(S_a) = D(S_b) * \mathbf{MD}(S_b)_{kk} = |\mathbf{M}_{S_aS_a}|$ (with $S_b = S_a \cup \{X_k\}$). This property can be easily derived from standard results on the determinants and inverses of partitioned matrices. Criteria not based on quadratic forms or determinants usually lead to computational difficulties that rend all-subset comparison procedures unfeasible for a moderate number of candidate variables.

VARIABLE SCREENING IN TWO-GROUP DISCRIMINANT ANALYSIS

Discriminant Analysis (DA) studies can have two major objectives: (i) Interpreting and describing group differences. (ii) Allocating entities to well-defined groups, based on a relevant set of entity attributes.

Techniques dealing with problems of the first type of are usually headed under the designation of descriptive DA, and methods dealing with problems of the second type have been called predictive DA, allocation or classification methods. However, the word classification is also used to designate Cluster Analysis methodologies and the expression predictive DA, may refer to Bayesian approaches to allocation problems (McLachlan 1992, pp 67-74). Therefore, in order to avoid unnecessary confusions, the term allocation will be adopted in this article. Descriptive DA techniques can also be used to interpret significant effects found by a multi-factor MANOVA or MANCOVA (Kobilinsky 1990, Masson 1990, Huberty 1994 pp 206). In this and the following sections, a more traditional view will be adopted, assuming that every group can be defined by the level of a single factor in a MANOVA model.

The choice of criteria to compare variable subsets in DA should pay particular attention to the objectives of the analysis. In particular, several methodologists (McKay and Campbel

1982a, 1982b, Huberty 1994) recommend that in descriptive studies, variables should be compared based on measures of group separation, while in allocation studies it is more appropriate to use estimates of prediction power. However, in two-group problems both approaches can lead to similar procedures. Consider the traditional approach to two-group allocation, based on the assumptions of multivariate normality and common within-groups variance-covariance matrix Σ. In that case, it is well known the optimal hit rate is a decreasing function of the Mahalanobis distance between group means, $\Delta = \left[(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \, \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)\right]^{1/2}$ (where $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ denote the means of groups 1 and 2). In a descriptive context, the most usual index of group separation is the sample analogue of Δ, $D = \left[(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)' \mathbf{S}^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)\right]^{1/2}$ (where $\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2$ and $\mathbf{S}$ are the sample mean vectors and sample pooled variance-covariance matrix). In allocation problems, D is also a reasonable criterion for subset comparisons, as several estimators of hit rates are based on it. For instance, assuming equal prior probabilities of group membership, Fisher proposed $\Phi(D/2)$ ($\Phi(.)$ being the cumulative probability of a standard normal variate) as an estimator for the actual hit rate of the classical allocation scheme, based on the Linear Discriminant Function (Fisher 1936). Although $\Phi(D/2)$ is optimistically biased, several authors (Lachenbruch 1968, Lachenbruch and Mickey 1968, McLachlan 1974) derived bias-correction factors, leading to parametric estimators that, once the sample dimension and number of variables are fixed, are simple functions of D.

The adaptation of Furnival algorithm to the comparisons of variable subsets based on D is straightforward. Following the notation of previous section, if $\mathbf{v}_{S_1}$ and $\mathbf{M}_{S_1 S_1}$ are chosen as $\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2$ and $\mathbf{S}$, then the conditions (A)-(B) are satisfied and $Q(S_a) = D^2_{S_a}$. Furthermore, D^2 never increases with the removal of variables, and the "leaps and bounds" approach of Furnival and Wilson can also be employed. Duarte Silva (forthcoming) has implemented this approach, using McLachlan's estimate of the hit rate (McLachlan 1974) as comparison criterion, $C(S_a)$. With the help of a personal computer, for several problems with 30 candidate variables, Duarte Silva was always able to identify to 20 best subsets according to $C(S_a)$ in less than 20 minutes of CPU time.

A potential drawback of the approach described above, is the fact that this approach is based on parametric estimators of hit rates, which relay on fairly strong assumptions and are not particularly robust (McLachlan 1986). Furthermore, there are many non-parametric estimators of hit rates that have good properties under a wide range of data conditions (McLachlan 1992, pp 337-366). These estimators typically require some "counting scheme" associated with a cross-validation strategy (ex: jackknife, leave-one-out or bootstrap) to avoid optimistic bias. As non-parametric estimates of hit rates are not based on quadratic forms or determinants, Furnival algorithm can not be easily adapted to their repeated computation.

Thus, the choice of criteria for subsets comparisons in allocation problems, may require evaluating a trade-off between stepwise procedures based on criteria (non-parametric estimates) with good properties for a wide range of data conditions, and all-subsets procedures based on criteria (parametric estimates) whose quality can only be guaranteed for specific conditions. A rigorous and complete analysis of this trade-off is beyond the scope of this article. However, given the article emphasis on all-subset comparisons as an alternative to stepwise procedures, an effort will be made in order to identify and characterize data conditions where the problems of stepwise methods are more pronounced.

Consider a two-group discriminant problem where the traditional assumptions of multivariate normality and equality of variance-covariances hold. Then, as discussed above, the population Mahalanobis distance associate with a subset S_a (Δ_{S_a}) is an appropriate measure of the S_a "quality" both for descriptive and allocation purposes. Thus, when an additional variable, X_k ($X_k \in \overline{S}_a$), is added to S_a leading to subset S_b, the contribution of X_k can be interpreted in terms of the increase in squared Mahalanobis distance, $\Delta^2_{S_b} - \Delta^2_{S_a}$. Using expressions (2)-(5). it follows that the increase in a quadratic form $Q(S_a)$ upon the addition of X_k is given by:

$$Q(S_b) - Q(S_a) = \frac{(\mathbf{v_k} - \mathbf{M_{k,S_a}} \mathbf{M^{-1}_{S_aS_a}} \mathbf{v_{S_a}})^2}{\mathbf{M_{kk}} - \mathbf{M_{k,S_a}} \mathbf{M^{-1}_{S_aS_a}} \mathbf{M_{S_a,k}}} \qquad (6)$$

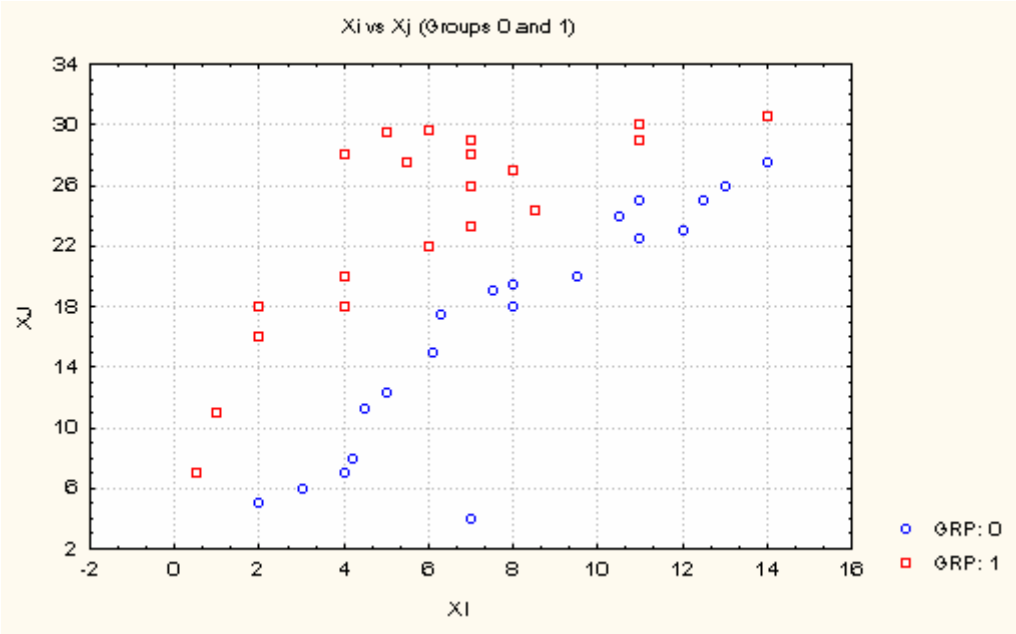
In particular, making $\mathbf{v_{S_1}} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ and $\mathbf{M_{S_1S_1}} = \boldsymbol{\Sigma}$, the increase in Δ^2 is given by:

$$\Delta^2_{S_b} - \Delta^2_{S_a} = \frac{\left[(\boldsymbol{\mu}_{1k} - \boldsymbol{\mu}_{2k}) - \boldsymbol{\Sigma}_{k,S_a} \boldsymbol{\Sigma}^{-1}_{S_aS_a} (\boldsymbol{\mu}_{1S_a} - \boldsymbol{\mu}_{2S_a})\right]^2}{\boldsymbol{\Sigma}_{kk} - \boldsymbol{\Sigma}_{k,S_a} \boldsymbol{\Sigma}^{-1}_{S_aS_a} \boldsymbol{\Sigma}_{S_a,k}} \qquad (7)$$

Formula (7) deserves close attention. It is known, from the properties of the multivariate normal distribution, that $E(X_k|S_a) = \alpha + \boldsymbol{\beta}\, \mathbf{X_{S_a}}$ with $\boldsymbol{\beta} = \boldsymbol{\Sigma}_{k,S_a} \boldsymbol{\Sigma}^{-1}_{S_aS_a}$. Thus, the numerator of (7) can be expressed as $\left[(\mu_{1k} - \mu_{2k}) - \beta(\mu_{1S_a} - \mu_{2S_a})\right]^2$. Therefore, the contribution of X_k to the group separation (as measured by Δ^2) is not directly related to its average difference across groups, $\boldsymbol{\mu}_{1k} - \boldsymbol{\mu}_{2k}$, but to the difference between $\boldsymbol{\mu}_{1k} - \boldsymbol{\mu}_{2k}$ and $\boldsymbol{\beta}(\boldsymbol{\mu}_{1S_a} - \boldsymbol{\mu}_{2S_a})$. This latter value can be interpreted as the across-group mean difference of X_k, explained by the relation between X_k and the elements of S_a. The denominator of (7) can be expressed alternatively as $\sigma^2_{X_k}(1 - \rho^2_{X_k.S_a})$, where $\sigma^2_{X_k}$ denotes the variance of X_k and $\rho^2_{X_k.S_a}$ the squared correlation between X_k and S_a. Thus, the denominator of (7) equals the conditional variance of X_k given S_a, and the contribution of X_k to Δ^2, is simply the ratio between the "non-explained" squared group-difference on X_k's average and X_k's conditional variance.

The result discussed above has important implications concerning the performance of stepwise procedures, as these procedures ignore the correlations between concurrent candidates to enter a model at a given step. Consider that in a forward stepwise procedure, two correlated variables, X_i and X_j, are currently kept out of the analysis. Assume, for the simplicity of the argument but without lack of generality, that X_i and X_j, are uncorrelated with the set of variables, S_a, already "in". In that case, the potential contribution of X_i and X_j, would be individually assessed by the ratios between their (sample) squared mean differences to their unconditional variances. If these ratios are small, X_i and X_j may never be included in the final analysis. However, when the (within-groups) correlation between X_i and X_j is strong, the conditional variance of X_j given X_i (or X_i given X_j) may be substantially overestimated by its unconditional counterpart, and any small unexpected difference of group means may have an important contribution to Δ^2. Figure 1 illustrates a typical situation where this phenomenon occurs.

Fig 1 - Illustration of data conditions unfavorable for forward stepwise selection methods (two-group DA)



It may be noticed that in this case the contribution of X_i and X_j to group separation, is mostly due to their different within-groups and across-groups correlation patterns, i.e., the fact that $\mu_{1i} - \mu_{2i}$ and $\mu_{1j} - \mu_{2j}$ have opposite signs, although within-groups X_i and X_j are positively correlated. Data conditions where these patterns differ are particularly unfavorable for forward stepwise procedures. However, backward stepwise procedures are not so strongly affected by this problem, because these procedures typically start the analysis by a model that includes all candidate variables. In the example described above, from any model including both X_i and X_j, the decrease in group separation due to the removal of one of these variables would be correctly assessed, and X_i or X_j would not be "good candidates" to leave the model.

There is however, another problem with stepwise procedures that may affect the performance of procedures based either on forward or backward strategies. This problem concerns the choice of the dimension of the subset(s) to be used in the final analysis. When the contributions to group separation are about evenly distributed among all variables, then any stepwise procedure, may fail to recognize when to stop (since all single variable increments or reductions in group separation will be similar) and suggest variable subsets that are either very small or very large.

It should be remarked that although through the discussion presented above, multivariate normality and equality of variance-covariances were assumed, the fundamental arguments presented should remain valid (although in some different form) in more general problems. In particular, differences in the within-groups and between-groups structure of variable dependence create serious problems for forward stepwise procedures, and individually small contributions to group differences by a large number of variables create difficulties for any stepwise (forward or backward) selection strategy.

VARIABLE SCREENING IN *K*-GROUP DISCRIMINANT ANALYSIS

Let's now turn our attention to variable screening for DA problems with more than two groups. In this case there is no single index of group separation with general applicability. Among the several indices proposed for descriptive purposes, one the most widely used is η^2, which may be defined as one minus the proportion of the sample generalized variance[(1)] that can not be explained by the group differences. Let *k* denote the number of groups, n_g the number of observations in group *g*, $N = \sum_{g=1}^{k} n_g$ the total number of observations, $\bar{\mathbf{x}}_\mathbf{g}$ and $\bar{\bar{\mathbf{x}}}$ the group *g*, and overall sample centroids, and $\mathbf{W} = (N\text{-}k)\mathbf{S}$, $\mathbf{B} = \sum_{g=1}^{k} n_g (\bar{\mathbf{x}}_\mathbf{g} - \bar{\bar{\mathbf{x}}})(\bar{\mathbf{x}}_\mathbf{g} - \bar{\bar{\mathbf{x}}})'$, $\mathbf{T} = \mathbf{W}+\mathbf{B}$ the within-groups, between-groups and total SSCP matrices of deviations from $\bar{\mathbf{x}}_\mathbf{g}$ (**W**) and $\bar{\bar{\mathbf{x}}}$ (**B** and **T)**. Then, $\eta^2 = 1 - |\mathbf{W}| / |\mathbf{T}| = 1 - \Lambda$, where Λ is the well known Wilk's statistic, concerning the null hypothesis of equal group means. McCabe (1975) has adapted Furnival algorithm to the comparison of variable subsets in DA, according to η^2 [(2)]. For that purpose the basic algorithm needs to be applied simultaneously to matrices **W** and **T**. After each symmetric sweep, the value of $\eta^2_{S_a}$ can be computed based on the updated determinants $|\mathbf{W}_{\mathbf{S_a}}|$ and $|\mathbf{T}_{\mathbf{S_a}}|$. As η^2 never increases with the removal of variables, the more efficient "leaps and bounds" approach of Furnival and Wilson can also be employed. Computer experiments show that replacing Furnival algorithm by Furnival and Wilson's can lead to impressive reductions in computational effort.

Another common index of group separation is the index *W* proposed by Rao (1952, pp 257), $W = \sum_{g=1}^{k} n_g (\bar{\mathbf{X}}_\mathbf{g} - \bar{\bar{\mathbf{X}}})' \mathbf{S}^{-1} (\bar{\mathbf{X}}_\mathbf{g} - \bar{\bar{\mathbf{X}}})$, which is an weighted sum of sample Mahalanobis distances between each group centroid and the overall centroid across all groups. The adaptation of Furnival and Wilson algorithm to subset comparisons based on this index offers no difficulties. We notice that W always increases with the addition of new variables and may be expressed as a sum of *k* quadratic forms $Q^{(g)}$**(.)** based on the vectors $\mathbf{v}^{(\mathbf{g})}_{\mathbf{S_1}} = \bar{\mathbf{X}}_\mathbf{g} - \bar{\bar{\mathbf{X}}}$ and common symmetric matrix $\mathbf{M}_{\mathbf{S_1 S_1}} = \mathbf{S}$**,** satisfying conditions (A) and (B). In this case, the *k* quadratic forms can be evaluated simultaneously if the usual **MV**(.) matrices are augmented with one raw and column for each $\mathbf{v}^{(\mathbf{g})}_{\mathbf{S_a}}$ [(3)].

Geisser (1977) and McCulloch (1986) studied the problem of measuring group separation, when the original data is projected into a space of dimension *q* ($q \le r = \min(p,k\text{-}1)$). These authors proposed a class of separatory measures consisting on all increasing functions of the *q* first eigenvalues of $(\mathbf{\Gamma}\text{-}\mathbf{\Sigma})\ \mathbf{\Sigma}^{-1}$ ($\mathbf{\Gamma}$ being the total variance-covariance of X). Replacing $\mathbf{\Gamma}$ and $\mathbf{\Sigma}$ by their sample estimates these measures lead to indices of separation based on the first *q* eigenvalues of $\mathbf{BW}^{-1}$, $\lambda_1, \lambda_2, ..., \lambda_q$. It can be shown that η^2 and *W* are two particular cases of such indices, as $\eta^2 = 1 - \prod_{i=1}^{r} \frac{1}{1+\lambda_j}$, $W = (N\text{-}k) \sum_{i=1}^{r} \lambda_i$, and η^2, *W* are indices that consider all possible dimensions of group separation. When all important group differences can be described by the first *q* linear discriminant functions, indices of group separation in the space generated by these functions, can be defined using versions of the usual indices, that ignore the last *r* - *q*

eigenvalues of $\mathbf{BW}^{-1}$. Alternatively, when it is particularly important to ensure that all dimensions of group-separation are correctly represented, a reasonable comparison criterion is λ_q, an index of the separation associated with the least dimension with importance. Unfortunately, Furnival algorithm can not be easily adapted to the repeated evaluation of λ_i (i=1,2,...,q). However, if $r \leq 3$ (typically problems with four or less groups) there is a way of computing the eigenvalues λ_i, from determinants and quadratic forms satisfying conditions (A)- (B), and all indices based on Geisser and McCulloch measures can be used as criteria in efficient all-subsets comparison procedures. Considering known relations between DA and Canonical Correlation Analysis (CCA), the λ_i can be presented as particular functions of canonical correlations. The adaptation of Furnival and Furnival and Wilson algorithms to the evaluation of λ_i, will be discussed latter, under the more general context of subset comparisons in CCA.

For DA problems involving more than two groups there is no simple relation between descriptive measures of group separation and estimates of prediction power. Schervish (1981) has proposed parametric estimators of hit rates, valid for the general *k*-group DA problem as long as the traditional assumptions hold. However, the computation of the resulting estimates requires the evaluation of *k*-1 dimensional integrals and Schervish estimators are not commonly used in practice. The estimation of prediction power in a *k*-group setting ($k > 2$) is usually based on non-parametric estimators that use some combination of "counting" and cross-validation strategies. As the computational burden required by both parametric and non-parametric estimators of hit rates is important, often it is not feasible to make all-subsets comparisons based on direct measures of prediction ability. Thus, in order to avoid the dangers of stepwise procedures one possibility is to use some index of group separation as a proxy for prediction ability. However, global indices of separation like η^2 or *W* are usually strongly influenced by the groups that are further apart, while prediction ability is mostly dependent on the separation between the groups that are closer together (McLachlan 1992, pp 93). An index that does not suffer from this problem is the smallest sample Mahalanobis distance between all pairs of groups. This index is used as a comparison criterion by one of the stepwise selection routines of SPSS. Furnival and Wilson algorithm can be easily adapted for subsets comparisons based on this criterion, by defining $\mathbf{M}_{\mathbf{S_1}\mathbf{S_1}}$ as $\mathbf{S}$ and creating C_2^k different $\mathbf{v}_{\mathbf{S_1}}$ vectors, one for each pair of differences in sample group centroids.

The characterization of the problems of stepwise procedures, presented previously in the context of two-group DA, generalizes naturally to *k*-group problems. In particular, the discussion presented applies directly to any measure based on Mahalanobis distances, like the population analogue of W or the distance between the two closest groups. If group separation is understood as a function of unexplained variance, measured by a population analogue of η^2, then some insight can be gained from following argument. Suppose that X_k is added to an analysis based on a subset S_a. Then the population counterpart of 1- η^2 is multiplied by a factor *c* (8) which is simply the ratio between the within-groups and total conditional variances of
X_k (given S_a).

$$c = \frac{\boldsymbol{\Sigma}_{kk} - \boldsymbol{\Sigma}_{k,S_a}\boldsymbol{\Sigma}_{S_aS_a}^{-1}\boldsymbol{\Sigma}_{S_a,k}}{\boldsymbol{\Gamma}_{kk} - \boldsymbol{\Gamma}_{k,S_a}\boldsymbol{\Gamma}_{S_aS_a}^{-1}\boldsymbol{\Gamma}_{S_a,k}} \qquad (8)$$

One of the principal problems of forward stepwise selection methods, is the fact that they ignore the correlations between variables currently out of the analysis. The consequences of this problem are bound to be more serious when these correlations have different impacts in the total and within-groups conditional variances. The occurrence of such different impacts is usually a consequence of differences in the within-groups and across-groups correlation structures. Finally, the problems of stepwise methods in finding an appropriate subset dimension, affect two-group and *k*-group DA problems in the same manner.

VARIABLE SCREENING IN MANOVA AND MANCOVA

Consider the general MANOVA (9) and MANCOVA (10) models:

$$\mathbf{Y} = \mathbf{X}\,\mathbf{\Pi} + \mathbf{U} \qquad (9)$$
$$\mathbf{Y} = \mathbf{X}\,\mathbf{\Pi} + \mathbf{Z}\,\mathbf{\Psi} + \mathbf{U} \qquad (10)$$

where **Y** is a (n*p) matrix of responses, **X** is a (n*q) design matrix, **Z** is a (n*t) matrix of covariates, **U** is a (n*p) matrix of error terms and $\mathbf{\Pi}$, $\mathbf{\Psi}$ are (q*p) and (t*p) matrices of unknown parameters. In this section, the problem of comparing the subsets, S_a, of **Y** according to their contribution to an "effect" characterized by the violation of a linear hypothesis H_0: $\mathbf{A}\,\mathbf{\Pi} = \mathbf{0}$, will be discussed.

General linear hypothesis $\mathbf{A}\,\mathbf{\Pi}\,\mathbf{C} = \mathbf{0}$, with **C** different of the (p*p) identity (**I**) will not be considered because of the following two reasons: (i) Usually, it only makes sense to select variable subsets concerning hypothesis that do not involve linear combinations of parameters associated with different variables. When such hypothesis are present (for instance, in the MANOVA approach to the analysis of repeated measurements) often all the response variables have substantive importance and none should be removed. (ii) Even when it makes sense to select subsets of **Y** concerning an hypothesis $\mathbf{A}\,\mathbf{\Pi}\,\mathbf{C} = \mathbf{0}$ ($\mathbf{C} \neq \mathbf{I}$) , usually it is not possible to find an appropriate criterion based on a matrix, **M,** satisfying condition (A) .

The usual test statistics concerning the hypothesis H_0 include Wilk's lambda ($\Lambda = |\mathbf{E}| / |\mathbf{T}|$), Bartlett-Pillai trace ($U = \mathrm{tr}\,\mathbf{H}\,\mathbf{T}^{-1}$) and Hotelling-Lawley trace ($V = \mathrm{tr}\,\mathbf{H}\,\mathbf{E}^{-1}$), where **H**, **E** and $\mathbf{T} = \mathbf{H} + \mathbf{E}$ are Hypothesis, Error and Total SSCP matrices (see e.g., Seber 1984, Ch. 9) concerning H_0. For the purpose of subset comparisons it is convenient to measure the extent of H_0 violations by an appropriate index of the magnitude of the associated effect. Several such indices may be defined, often as a function of the rank of the matrix **H** (*r),* and the value of some test statistic. For instance, three usual indices of effect magnitude are $\tau^2 = 1 - \Lambda^{1/r}$, $\xi^2 = U/r$ and $\varsigma^2 = V / (V+r)$.

It may be noticed that Descriptive DA can be presented within this framework. For instance, the problems covered in the previous sections are related to a MANOVA model (9) where the design matrix **X** implies a one-way layout. In that case the hypothesis H_0 concerns the equality of group means and the associated effect may be described as "group separation". The corresponding SSCP matrices are respectively $\mathbf{E} = \mathbf{W}$ and $\mathbf{H} = \mathbf{B}$ and the rank of **H** equals the minimum between the number of responses and number of groups minus one ($r = \min(p, k-1)$). However, the discussion to be presented in this section is more general, because on the one hand it allows for the presence of covari-

ates, and on the other hand it admits subset comparisons according to their contribution to other effects, like specific contrasts or factor interactions in a multi-way layout.

Furnival and Wilson algorithm can be adapted to subset comparisons based on any criterion that can be expressed as a function of the value of the statistics, Λ, U or V. First, notice that for an hypothesis $\mathbf{A\,\Pi} = \mathbf{0}$, all three matrices **H**, **E** and **T** satisfy condition (A). Furthermore, when a variable Y_k is removed from a subset S_b, the value of Λ always increases and the values of U and V always decrease. Subset comparisons based on Λ (or more appropriately on τ^2) follow from a direct application of McCabe (1975) strategy to the matrices **E** and **T,** replacing Furnival algorithm by Furnival and Wilson's. To show that this algorithm can also be applied to subset comparisons based on criteria derived from U or V, it suffices to show that these statistics can be expressed as sums of quadratic forms $Q^{(i)}(.)$. Let $\sum_{i=1}^{r} \theta_i \mathbf{h_i}\,\mathbf{h_i}'$ be the spectral decomposition of **H**. Then, U can be expressed alternatively as U = tr $\mathbf{H\,T^{-1}} = tr\left[\left(\sum_{i=1}^{r}\theta_i\ \mathbf{h_i}\ \mathbf{h_i}'\right)\mathbf{T}^{-1}\right] = \sum_{i=1}^{r}\left(\sqrt{\theta_i}\mathbf{h_i}\right)'\mathbf{T^{-1}}\left(\sqrt{\theta_i}\mathbf{h_i}\right)$ where θ_i is the i-th eigenvalue of **H** and $\mathbf{h_i}$ the corresponding normalized eigenvector. Thus, the adaptation of Furnival and Wilson algorithm simply requires an evaluation of the sums $\sum_{i=1}^{r} Q^{(i)}(.)$ based on the matrix $\mathbf{M_{S_1S_1}} = \mathbf{T}$, and vectors $\mathbf{v_{S_1}^{(i)}} = \sqrt{\theta_i}\,\mathbf{h_i}$,. Using a similar argument, subset comparisons based on criteria derived from V = tr $\mathbf{H\,E^{-1}}$ can be easily made if $\mathbf{M_{S_1S_1}}$ and $\mathbf{v_{S_1}^{(i)}}$ are defined as **E** and $\sqrt{\theta_i}\,\mathbf{h_i}$.

Stepwise selection methods in MANOVA and MANCOVA are usually based on Roy's additional information criterion, $\Lambda_{k|S_a}$ (11), (Rao 1973) which measures the contribution of Y_k to the effect under study, when the influence of the variables included in S_a is factored out.

$$\Lambda_{k.S_a} = \frac{\mathbf{E_{kk}} - \mathbf{E_{k,S_a}}\mathbf{E_{S_aS_a}^{-1}}\mathbf{E_{S_a,k}}}{\mathbf{T_{kk}} - \mathbf{T_{k,S_a}}\mathbf{T_{S_aS_a}^{-1}}\mathbf{T_{S_a,k}}} \qquad (11)$$

The denominator of (11) is proportional to the sample conditional variance of Y_k, and the numerator can be interpreted as the portion of Y_k 's sample conditional inertia that is not related to the effect. Considering the population equivalents to **E** and **T**, it can be argued that forward stepwise methods are more prune to miss good subsets when there are important differences between the correlation structures related and unrelated to the effect under study. On the other hand, both forward and backward stepwise procedures may have problems identifying appropriate subset dimensions, particularly when the effect contributions are about evenly distributed by a large number of different variables.

VARIABLE SCREENING IN CANONICAL CORRELATION ANALYSIS

Canonical Correlation Analysis (CCA) is traditionally described as a technique to study the relations between two sets of variables. This view of CCA has interest in applications where two different concepts are measured by several variables, for example, a medical researcher

studying the relation between sets of variables describing eating habits and heart conditions or a financial analyst trying to relate ratios of capital structure to measures of profitability. However, the role of CCA in multivariate statistics is not restricted to its use as a multivariate technique on its own. The theory of CCA is particularly important because it provides a unifying framework for several multivariate methodologies. If effect, Discriminant Analysis, Multivariate Regression Analysis, MANOVA and MANCOVA can all be presented as particular cases of CCA.

In the first part of this section, variable screening will be discussed within the context of CCA as a technique on its own. In the second part, the relations between CCA and other methodologies will be explored, showing how the results presented in the first part relate to and sometimes extend, the methods of variable screening discussed in the previous sections.

Let $X = \{X_1, X_2, ..., X_q\}$ and $Y = \{Y_1, Y_2, ..., Y_p\}$ be two sets of variables and denote the dimension of the smaller set by $r = \min(q, p)$. Suppose that one of these sets, say X, is fixed and it is desired to compare the subsets, S_a, of the other set, Y, according to some measure of association between S_a and X. Several measures of multivariate association have been proposed in the literature. Cramer and Nicewander (1979) argue that good measures of association should be invariant to linear transformations, and symmetric, i.e., should not be affected by a reversal in the roles played by the X and Y sets. These authors discuss seven alternative measures with those properties, all of which are functions of sample squared canonical correlations ($\hat{\rho}_i^2$) between X and Y.

The measures considered by Cramer and Nicewander are the following:

$$\hat{\rho}_1^2, \quad \hat{\gamma}_1 = \prod_{i=1}^{r} \hat{\rho}_i^2 \quad , \quad \hat{\gamma}_2 = 1 - \prod_{i=1}^{r} (1 - \hat{\rho}_i^2) \quad ; \quad \hat{\gamma}_3 = 1 - \frac{r}{\sum_{i=1}^{r} 1/(1 - \hat{\rho}_i^2)} \quad ; \quad \hat{\gamma}_4 = \left[\prod_{i=1}^{r} \hat{\rho}_i^2 \right]^{1/r} \quad ;$$

$$\hat{\gamma}_5 = 1 - \left[\prod_{i=1}^{r} (1 - \hat{\rho}_i^2) \right]^{1/r} \; ; \; \hat{\gamma}_6 = \frac{\sum_{i=1}^{r} \hat{\rho}_i^2}{r}$$

For the purpose of subset comparisons these measures result in five alternative criteria as $\hat{\gamma}_1$, $\hat{\gamma}_4$ and $\hat{\gamma}_2$, $\hat{\gamma}_5$ are monotically related ($\hat{\gamma}_4 = \hat{\gamma}_1^{1/r}$; $\hat{\gamma}_5 = 1 - (1 - \hat{\gamma}_2)^{1/r}$).

The first squared canonical correlation, $\hat{\rho}_1^2$, measures the linear association between the maximally correlated linear combinations of X and Y. Therefore, $\hat{\rho}_1^2$ is an appropriate measure of multivariate association, when it reasonable to assume that each set of variables can be summarized along one single dimension. All the other measures consider the *r* possible dimensions along which X and Y may be associated. The measure $\hat{\gamma}_1$ was initially proposed (in slightly different contexts) by Hotelling (1936) and Cramer (1974). Its use can be justified by the following argument. Suppose that the values of the *r* variables belonging to the smaller set, are predicted by linear regressions on the variables belonging to the larger set. Then, it can be shown that the ratio between the generalized variances of the predicted and observed values equals $\hat{\gamma}_1$. In that sense, $\hat{\gamma}_1$ can be interpreted as a multivariate generalization of the coefficient of determination, R^2. The use of $\hat{\gamma}_4$ is justified by noting that while a (univariate) variance can be described in terms of a vector's squared-length, a generalized variance has an equivalent interpretation in terms of the squared-volume of a parallelotope (Anderson 1958). Thus, replacing $\hat{\gamma}_1$ by $\hat{\gamma}_4$ is equivalent to replacing the ratio

between two squared-volumes by the ratio between the side squared-lengths of the hyper-cubes with the same volume as the original parallelotopes (Cramer and Nicewander 1979, pp 49). The measure $\hat{\gamma}_2$ was proposed by Hotelling (1936) and Rozebom (1965), and can be justified along similar lines to $\hat{\gamma}_1$. In effect, $\hat{\gamma}_2$ is also a natural generalization of R^2 for the regressions of the variables in the smaller set on the variables of the larger set. It may be shown that in this case, $\hat{\gamma}_2$ equals one minus the ratio between the generalized variances of the residuals and the observed values. However, $\hat{\gamma}_2$ is not equal to $\hat{\gamma}_1$ because in multivariate regression the generalized variances of the residuals and predicted values do not add up to the generalized variance of the observed values. The use of $\hat{\gamma}_5$ instead of $\hat{\gamma}_2$ follows from the same argument that justifies replacing $\hat{\gamma}_1$ by $\hat{\gamma}_4$. The measure $\hat{\gamma}_3$ was proposed independently by Coxhead (1974) and Shaffer and Gillo (1974). Assuming the same set of linear regressions as previously, it can be shown that the ratio between the sums of the predicted and observed squared-distances (according to a Mahalanobis metric) between all pairs of observations, equals $\hat{\gamma}_3$. The measure $\hat{\gamma}_6$ is the average of the squared canonical correlations and is the measure preferred by Crammer and Nicewander because, among other reasons, of its simplicity. Additional desirable properties of $\hat{\gamma}_6$ are discussed in Crammer and Nicewander (1979) article.

In order to adapt Furnival and Wilson algorithm to subset comparisons based on the measures described above, we first note that none of these measures can increase when variables are removed. Furthermore, it can be shown that if the Y variables are regressed on the X variables, then $\hat{\rho}_i^2$ equals the i-th principal eigenvalue of $\mathbf{S}_{\hat{\mathbf{Y}}\hat{\mathbf{Y}}}\mathbf{S}_{\mathbf{YY}}^{-1}$ where $\mathbf{S}_{\hat{\mathbf{Y}}\hat{\mathbf{Y}}}$ and $\mathbf{S_{YY}}$ are the sample variance-covariance matrices of the predicted and observed Y values (conversely, when the X's are regressed on the Y's, $\hat{\rho}_i^2$ equals the i-th eigenvalue of $\mathbf{S}_{\hat{\mathbf{X}}\hat{\mathbf{X}}}\mathbf{S}_{\mathbf{XX}}^{-1}$). Therefore, noting that the i-th eigenvalues of $\mathbf{S}_{\mathbf{ee}}\mathbf{S}_{\mathbf{YY}}^{-1}$ and $\mathbf{S}_{\hat{\mathbf{Y}}\hat{\mathbf{Y}}}\mathbf{S}_{\mathbf{YY}}^{-1}$ (where $\mathbf{S_{ee}}$ denotes the variance-covariance matrix of the residuals for the regressions of Y on X) add up to one, $\hat{\gamma}_2$ may be expressed as $\hat{\gamma}_2 = 1 - |\mathbf{S_{ee}}| / |\mathbf{S_{yy}}|$. Thus, efficient subset comparisons based on $\hat{\gamma}_2$ (or $\hat{\gamma}_5$) only require applying McCabe strategy simultaneously to the matrices $\mathbf{S_{ee}}$ and $\mathbf{S_{yy}}$. Comparisons based on $\hat{\gamma}_3$, follow from $\hat{\gamma}_3 = \text{tr}\,(\mathbf{S}_{\hat{\mathbf{Y}}\hat{\mathbf{Y}}}\mathbf{S}_{\mathbf{ee}}^{-1}) / \text{tr}\,(\mathbf{S}_{\mathbf{YY}}\mathbf{S}_{\mathbf{ee}}^{-1}) = \text{tr}(\mathbf{S}_{\hat{\mathbf{Y}}\hat{\mathbf{Y}}}\mathbf{S}_{\mathbf{ee}}^{-1}) / [r + \text{tr}\,(\mathbf{S}_{\hat{\mathbf{Y}}\hat{\mathbf{Y}}}\mathbf{S}_{\mathbf{ee}}^{-1})]$. Using the spectral decomposition of $\mathbf{S}_{\hat{\mathbf{Y}}\hat{\mathbf{Y}}}$, $\text{tr}(\mathbf{S}_{\hat{\mathbf{Y}}\hat{\mathbf{Y}}}\mathbf{S}_{\mathbf{ee}}^{-1})$ can be expressed as a sum of r quadratic forms $\mathbf{v}_{\mathbf{S}_1}^{(\mathbf{i})\prime}\mathbf{M}_{\mathbf{S}_1\mathbf{S}_1}^{-1}\mathbf{v}_{\mathbf{S}_1}^{(\mathbf{i})}$, where $\mathbf{M}_{\mathbf{S}_1\mathbf{S}_1}$ equals $\mathbf{S_{ee}}$ and $\mathbf{v}_{\mathbf{S}_1}^{(\mathbf{i})}$ is an eigenvector of $\mathbf{S}_{\hat{\mathbf{Y}}\hat{\mathbf{Y}}}$. Then, the adaptation of Furnival and Wilson algorithm is straightforward. Efficient subset comparisons based on $\hat{\gamma}_6$ can also be made without any major difficulty. Noting that $\hat{\gamma}_6 = \text{tr}\,(\mathbf{S}_{\hat{\mathbf{Y}}\hat{\mathbf{Y}}}\mathbf{S}_{\mathbf{YY}}^{-1}) / r$, and using the spectral decomposition of $\mathbf{S}_{\hat{\mathbf{Y}}\hat{\mathbf{Y}}}$, $r\hat{\boldsymbol{\gamma}}_6$ can also be expressed as a sum of quadratic forms, $Q^{(i)}(.)$, and the usual procedure follows.

Efficient subset comparisons based on $\hat{\gamma}_1$ ($\hat{\gamma}_4$) or $\hat{\rho}_1^2$ are not as straightforward, as those based on the other measures. If $p = \min(p,q) = r$, i.e., if the number of variables under comparison (Y) does not exceed the dimension of the fixed set (X), then $\hat{\gamma}_1 = |\mathbf{S}_{\hat{\mathbf{Y}}\hat{\mathbf{Y}}}| / |\mathbf{S_{YY}}|$ and subset comparisons based on $\hat{\gamma}_1$ (or $\hat{\gamma}_4$) can be made by a simple adaptation of McCabe (1975) algorithm. However, in most applications p is larger than q, which implies that $|\mathbf{S}_{\hat{\mathbf{Y}}\hat{\mathbf{Y}}}|$

= 0, and $\hat{\gamma}_1$ can not be easily expressed as a ratio of determinants. As far as we can tell, there is no general way of adapting Furnival and Furnival and Wilson algorithms to subset comparisons based on $\hat{\gamma}_1$,$\hat{\gamma}_4$ or $\hat{\rho}_1^2$. However, in some special cases it is possible to take advantage of known relations between traces and determinants, in order to compute all (or some) canonical correlations for the subsets under comparison. For instance, if $r \leq 3$ the known relations $\Lambda = \prod_{i=1}^{r}(1-\hat{\rho}_i^2) = |\mathbf{S_{ee}}| \ / \ |\mathbf{S_{yy}}|$, $U = \sum_{i=1}^{r}\hat{\rho}_i^2 = \text{tr}(\mathbf{S}_{\hat{\mathbf{Y}}\hat{\mathbf{Y}}}\mathbf{S}_{\mathbf{YY}}^{-1})$, $V = \sum_{i=1}^{r}\left(\hat{\rho}_i^2/(\mathbf{1}-\hat{\rho}_i^2)\right) = \text{tr}(\mathbf{S}_{\hat{\mathbf{Y}}\hat{\mathbf{Y}}}\mathbf{S}_{\mathbf{ee}}^{-1})$, (where Λ, U and V denote the values of Wilk's, Bartlett-Pillai and Hotelling-Lawley statistics concerning the null hypothesis of no association between the X and Y sets), can be used to compute all $\hat{\rho}_i^2$. After some tedious algebra it follows that for r = 2, the values of $\hat{\rho}_1^2$ and $\hat{\rho}_2^2$ are given by equation (12) and for r = 3, $\hat{\rho}_1^2, \hat{\rho}_2^2$ and $\hat{\rho}_3^2$ are the three solutions of equation (13).

$$r = 2 \ \Rightarrow \ \hat{\rho}_i^2 = \frac{1}{2}\left(U \pm \sqrt{U^2 - 4(U + \Lambda - 1)}\right) \qquad (12)$$

$$r = 3 \ \Rightarrow \ (\hat{\rho}_1^2)^3 - U(\hat{\rho}_1^2)^2 + [2U\text{-}3 + \Lambda(V+3)](\hat{\rho}_1^2) + 2\text{-}U\text{-}\Lambda(V+2) = 0 \qquad (13)$$

The results presented above, have important implications in other multivariate techniques rather than CCA on its own. For instance, it is well known that *k*-group descriptive DA can be presented has the CCA between *p* variables describing entity attributes, and *k-1* indicator variables describing group membership (McLachlan 1992, pp 185-187). In that case, $\mathbf{S}_{\hat{\mathbf{Y}}\hat{\mathbf{Y}}}\mathbf{S}_{\mathbf{ee}}^{-1}$ and $\mathbf{BW^{-1}}$ have the same positive eigenvalues (λ_i) and the measure $\hat{\gamma}_2$ is the usual η^2 index. Furthermore the squared canonical correlations are related to the eigenvalues of $\mathbf{BW^{-1}}$ by the equation $\lambda_i = \hat{\rho}_i^2/\left(1-\hat{\rho}_i^2\right)$.

Consider now the MANOVA (9) and MANCOVA (10) models. Denote the space spanned by the columns of X (MANOVA) or X and Z (MANCOVA) by Ω, the subspace of Ω defined by the hypothesis H_0: $\mathbf{A\,\Pi} = \mathbf{0}$ by ω, the orthogonal complement of ω by $\omega^{\perp}$ and the space spanned by the projection of Y on $\omega^{\perp}$ by γ. Then, the analysis of H_0 can be presented in terms of a CCA between two sets of vectors spanning γ and Ω (Masson 1990). In this case, $\mathbf{S}_{\hat{\mathbf{Y}}\hat{\mathbf{Y}}}\mathbf{S}_{\mathbf{ee}}^{-1}$ and $\mathbf{HE^{-1}}$ have the same positive eigenvalues, the $\hat{\gamma}_3$,$\hat{\gamma}_5$, $\hat{\gamma}_6$ measures are the usual ς^2, τ^2 and ξ^2 indices, and the rank of **H**, *r*, equals the minimum between *p* and the dimension of
$\omega^p = \omega^{\perp} \cap \Omega$. When $r \leq 3$, Furnival algorithm can be adapted for subset comparisons based on any function of the *r* non-zero eigenvalues of $\mathbf{HE^{-1}}$ (λ_i), or $\mathbf{HT^{-1}}$ ($\hat{\rho}_i^2$), since if r=1, $\hat{\rho}_1^2$ = U = tr $\mathbf{HT^{-1}}$, λ_1 = V = tr $\mathbf{HE^{-1}}$, and if $1 < r \leq 3$ equations (12) or (13) hold.

COMPUTATIONAL EFFORT

The evaluation of the computational effort required by all-subsets comparison algorithms is usually based on the number floating point operations performed. In the case of the algorithms discussed in this article the following three questions are particularly relevant: (i) What is the effort required by algorithms based on Furnival's (or McCabe's) exhaustive evaluation of all quadratic forms *Q*(.) and/or determinants *D*(.) ? (ii) What are the computa-

tional savings when an exhaustive search is replaced by Furnival and Wilson implicit enumeration algorithm ? (iii) What is the effort required to convert quadratic forms and determinants to appropriate comparison criteria ?

For question (i) exact answers can be found. In effect, as discussed in section 2, Furnival algorithm consists essentially on a succession of symmetric sweeps on portions of *(p+1)*(p+1)* matrices, **MV**(.), or *p*p* matrices **MD**(.) with 2^{p-t-1} different sweeps involving *t* additional variables (*t = 0,...,p-1).* As the number of floating point operations required by each sweep is a quadratic function on *t*, the evaluation of the computational effort for the different versions of Furnivals's algorithm can be made with the help of the known results on the summations $\sum_{t=0}^{p-1} 2^{-t}$, $\sum_{t=0}^{p-1} t\,2^{-t}$, $\sum_{t=0}^{p-1} t^2\,2^{-t}$ presented in equations (14)-(16).

$$\sum_{t=0}^{p-1} 2^{-t} = 2 - (1/2)^{p-1} \tag{14}$$

$$\sum_{t=0}^{p-1} t\ 2^{-t} = 2 - (1/2)^{p-1}\,(p+1) \tag{15}$$

$$\sum_{t=0}^{p-1} t^2\ 2^{-t} = 6 - (1/2)^{p-1}\ p^2 - (1/2)^{p-2}\ p - 3(1/2)^{p-1} \tag{16}$$

In particular, for the evaluation of quadratic forms, each sweep requires *(t+1)(t+4)/2 = 1/2* (t^2 *+ 5t + 4)* operations, and after some trivial computations, it follows that the total number of floating point operations required is of the order $6(2^p)$. This result is relevant for subset comparisons in linear regression, and two-group DA comparisons based on D^2 or on parametric hit rate estimates. Subset comparisons in *k*-group DA, MANOVA, MANCOVA or CCA based on any function of Bartlett-Pillai *U* or Hotelling-Lawley *V* require the evaluation of *r* quadratic forms with the same symmetric matrix. In that case each sweep requires *t (t+3)/2+ (t+2) r* = 1/2 [t^2 + (3 + 2*r*) *t* + 4r] operations, and the total number of operations is of the order $(3+3r)(2^p)$. Subset comparisons based on the smallest Mahalanobis distance in *k*-group DA, require the evaluation of C_2^k quadratic forms. In this case each sweep requires $t(t+3)/2 + (t+2)C_2^k = 1/2\left[t^2 + \left(3 + 2C_2^k\right)t + 4C_2^k\right]$ operations, and the total number of operations is of the order $\left(3+3C_2^k\right)\left(2^p\right)$. Subset comparisons based on functions of Wilk's Λ (like the η^2, τ^2 indices in *k*-group DA, MANOVA or MANCOVA, or equivalently the $\hat{\gamma}_2$ or $\hat{\gamma}_5$ measures in CCA) require the evaluation of all determinants *D*(.), for two different sets of matrices **MD**(.). Each sweep in MacCabe's adaptation of Furnival's algorithm requires *t(t+3)/2 + 1 = 1/2 (*t^2*+3t+2)* operations, and the total number of operations for each set of matrices is of the order $4(2^p)$. Thus, the effort required for the evaluation of Λ for all variable subsets is of the order $8(2^p)$. In three-group DA, MANOVA, MANCOVA or CCA with *r*=2, subset comparisons based on $\hat{\rho}_1^2$ or $\hat{\rho}_2^2$ (λ_1 or λ_2), can be made using equation (12), which requires finding the values of Λ and U. Noticing that the repeated evaluation of these two statistics require sweeps on the same matrix (**T** or $\mathbf{S_{YY}}$) some computation savings can be achieved. In effect, after sweeping the **MD**(.) matrices used in the evaluation of Λ_{S_a} , only the *r* last rows of the **MV**(.) matrix associated with U_{S_a} need to be updated. Therefore, each sweep in the simultaneous evaluation of Λ and U requires *t* (*t*+3+*r*) + 2 (1 + *r*) = t^2 + 5 *t* + 6 operations and the total number of operations is of the order $14(2^p)$. Finally, subset compari-

sons based on any function of $\hat{\rho}_1^2$, $\hat{\rho}_2^2$, $\hat{\rho}_3^2$ (λ_1, λ_2 , λ_3) in four-group DA, MANOVA, MANCOVA or CCA with $r = 3$, can be made after solving equation (13), which requires the simultaneous evaluation of Λ, U and V. Noting that the two matrices required for computing all Λ_{S_a}, are exactly the same matrices used in the evaluation of U_{S_a} and V_{S_a}, it follows that in this case each sweep requires $t\ (t+3+2r) + 2(1+2r) = t^2 + 9\ t + 14$ operations and the total number of operations is of the order $26(2^p)$. Table 1 presents the exact number of floating operations required by all the versions of Furnival algorithm discussed in this article.

TABLE 1: Computational effort required by exhaustive comparison procedures based on Furnival algorithm

	Number of floating point operations
$Q(.)$	$6(2^p) - (1/2)\ p^2 - (7/2)\ p - 6$
$D(.)$	$4(2^p) - (1/2)\ p^2 - (5/2)\ p - 4$
Λ	$8(2^p) - p^2 - 5\ p - 8$
$U \vee V$	$(3 + 3\ r)\ (2^p) - (1/2)\ p^2 - (5/2 + r)\ p - (3 + 3\ r)$
$\Lambda \wedge U$ (r =2)	$14(2^p) - p^2 - 7\ p - 14$
$\Lambda \wedge U \wedge V$ (r=3)	$26(2^p) - p^2 - 11\ p - 26$
Min (D^2)	$(3 + 3\,C_2^k)\ (2^p) - (1/2)\ p^2 - (5/2 + C_2^k)\ p - (3 + 3\,C_2^k)$

Question (ii) can not be answered exactly because the portion of Furnival and Wilson's search tree that can be pruned is dependent on the configuration of the sample data. In particular, when different sets of variables have highly different impacts over the comparison criterion, $C(.)$, it is relatively easy to identify "the best" subsets early on, and large parcels of the search tree can be pruned. On the other hand, when all the variables have similar contributions to $C(.)$, it is more difficult to eliminate subsets and the computational savings tend to be smaller. In spite of this difficulty, a rough idea of the computational burden for some typical and worst case scenarios can be given based on simulation experiments. For instance, Furnival and Wilson (1974) report that in a series of trials to find the 10 best subsets of each dimension, in regression models, the number of operations performed was equal to 3,764 (p=10), 123,412 (p=20), 3,934,714 (p=30) and 11,614,024 (p=35). These figures correspond to 62.18%, 1.19%, 0.06% and 0.005% of the number of operations that would be required to compare all subsets using Furnival algorithm. Duarte Silva (forthcoming) in a series of trials to identify the 20 best subsets according to McLachlan (1974) hit rate estimate in two-group DA, reports a number of operations in the range 3,619 - 4,494 (p=10), 117,143 - 557,660 (p=20) and 7,462,258 - 57,380,009 (p=30) which correspond respectively to 61% - 74% (p=10), 1.86% - 8.86% (p=20), and 0.15% - 0.89% (p=30) of the effort that would be required by Furnival algorithm. Duarte Silva experiments refer to worst case scenarios where all variables were generated from populations where these variables had equal discriminatory power. A few trials with other procedures revealed efforts of similar orders of magnitude. In particular, for $p \geq 30$, comparison procedures based on Furnival and Wilson algorithm are

typically faster than procedures based on Furnival algorithm, at least by a factor of 100, and as *p* grows this factor quickly becomes larger than 1000.

All the analysis presented in the previous paragraphs considered the number of operations involved in the computation of quadratic forms, *Q*(.), and/or determinants, *D*(.), and ignored the effort required to convert *Q*(.) and *D*(.) into appropriate comparison criteria, *C*(.). However, using efficient implementations, it is usually possible to minimize the number of conversions, ensuring that they require a small fraction of the total computational effort.

For instance, when *C*(.) is equal to *Q*(.) (like the squared Mahalanobis distance in two-group DA, or the smallest squared Mahalanobis distance in *k*-group DA) or is a monotic function of *Q*(.) summations (like Roy's *W* index in *k*-group DA, the ξ^2 , ς^2 indices in MANOVA/MANCOVA or the $\hat{\gamma}_3$, $\hat{\gamma}_6$ measures in CCA), subset comparisons can be made directly on *Q*(.), and no conversion is required for that purpose. Even if the final results are presented in terms of *C*(.), conversions are required only for the pool of subsets selected for further analysis, which for pratical reasons should never include more than a few hundred subsets.

In two-group DA problems, for subsets of equal dimensions, comparisons based on parametric estimates of the hit rate are equivalent to comparisons based on the sample Mahalanobis distance and do not require any conversion from *Q*(.) to *C*(.). However, that is no longer the case for comparisons across subsets of different dimensions. In that case, the number of conversions can be minimized if for each subset, S_a, $Q(S_a) = D^2_{S_a}$ is compared against the largest squared Mahalanobis distance for the subsets of the same dimension, already excluded from the list of candidates for further analysis. Only when $Q(S_a)$ exceeds this value, a conversion from $Q(S_a)$ to $C(S_a)$ is required. Some computational experiments show that for values of *p* around 30, while the evaluation of the *Q*(.) typically require a few million operations, the number of conversions usually does not exceed a few hundred. Thus, even when relatively expensive estimates are chosen for *C*(.), these conversions still remain a small fraction of the total effort.

Subset comparisons based on monotic functions of Wilk's Λ statistic can be made using the same the number of operations as those required to compute the determinants $|\mathbf{E}_{S_a}|$ and $|\mathbf{T}_{S_a}|$. In effect, if $\mathbf{MD}^{(1)}(.)$ and $\mathbf{MD}^{(2)}(.)$ are the matrices derived respectively from **E** and **T**, in McCabe's adaptation of Furnival's algorithm, at each sweep the value of Λ can be updated directly using the relation $\Lambda_{S_b} = \Lambda_{S_a} * (\mathbf{MD}^{(1)}(S_a)_{kk} / \mathbf{MD}^{(2)}/ (S_a)_{kk})$ which only requires two operations.

In DA, MANOVA/MANCOVA and CCA problems with *r* = 2, subset comparisons based on a arbitrary function of $\hat{\rho}_1^2$ and $\hat{\rho}_2^2$ require solving equation (12) which involves three operations and the extraction of a square root. However, for many special cases this effort can be reduced. For instance, if *C*(.) is a monotic function of $\hat{\rho}_1^2$, then the computation of $\hat{\rho}_2^2$ can be skipped which saves one floating point operation. Furthermore, the relation $\left(U_{S_i} > U_{S_j}\right) \wedge \left(\mathbf{\Lambda}_{S_i} < \mathbf{\Lambda}_{S_j}\right) \Rightarrow \hat{\rho}^2_{1S_i} > \hat{\rho}^2_{1S_j}$ can be used in order to minimize the number of conversions. Conversely, if *C*(.) is a monotic function of $\hat{\rho}_2^2$, the number of conversions can be reduced noting that $\left(U_{S_i} > U_{S_j}\right) \wedge \left(1/4(U^2_{S_i} - U^2_{S_j}) + (\mathbf{\Lambda}_{S_j} - \mathbf{\Lambda}_{S_i}) < U_{S_i} - U_{S_j}\right) \Rightarrow \hat{\rho}^2_{2S_i} > \hat{\rho}^2_{2S_j}$.

The evaluation of $\hat{\rho}_1^2, \hat{\rho}_2^2$ and $\hat{\rho}_3^2$ in DA, MANOVA/MANCOVA and CCA problems with *r* = 3, involves solving the third-degree equation (13). In our current implementation, equation (13) is solved by the Newton-Raphson method. However, for common *C*(.), based on the

behavior of the function defined by the left hand side of (13), and noting that $0 \leq \hat{\rho}_3^2 \leq \hat{\rho}_2^2 \leq \hat{\rho}_1^2 \leq 1$, it is often possible to avoid unnecessary conversions.

The principal justification for the analysis of computational effort presented in this section, is to have an idea the number of candidate variables, *p* for which the procedures presented in this article, are feasible within a "reasonable" time. Two different views of "reasonable" will be adopted. In the first view, directed to on-line analysis, the computational time will be considered as "reasonable" if it does not exceed three minutes. In the second view, directed to batch analysis, the computational time will be considered as "reasonable" if it does not exceed 48 hours (what might be called a "week-end analysis"). In modern personal computers, the number of floating point operation performed per second typically varies between one hundred thousand and one million. Thus, the limit on the number of operations for on-line all subset comparisons should lie somewhere in the range 18 - 180 millions. Assuming that for the largest on-line analysis, Furnival and Wilson algorithm is about 1000 times faster than Furnival's, and ignoring the overhead and conversion effort (which for problems of this size is reasonable), the maximum number of candidate variables that can be analyzed on-line, should be in the range 30-35. "Week-end" batch analysis can be performed as long as the number of operations required is not above 17-170 billions. Assuming that for the largest problems, Furnival and Wilson algorithm is about 100,000 times faster than Furnival's, the maximum allowable value for *p* should be around 45-50. Obviously, these figures should be interpreted only as rough estimates, since the true limit on *p* depends on many factors, such as the type of computer used, analysis performed, criteria chosen and data configuration. When this limit is exceeded, a viable alternative, is to use either judgment, substantive knowledge, or stepwise selection methods (if the data conditions are not particularly unfavorable) in order to reduce the number of candidate variables to a manageable size, and then employ an all-subsets comparison procedure.

REFERENCES

Anderson, T.W. 1958. An Introduction to Multivariate Statistical Analysis. John Wiley. New York, NY.

Beale, E.M.L., Kendall, M.G. and Mann, D.W. 1967. The discarding of variables in multivariate analysis. Biometrika. 54: 357-366.

Coxhead, P. 1974. Measuring the relationship between two sets of variables. *British* Journal of Mathematical and Statistical Psychology. 27: 205-212.

Cramer, E.M. 1974. A generalization of vector correlation and its relation to canonical correlation. Multivariate Behavioral Research. 9: 347-352.

Cramer, E.M. and Nivewander, W.A. 1979. Some symmetric invariant measures of multivariate association. Psychometrika. 44 (1), 43-54.

Derksen, S. and Keselman, H. 1992. Backward, forward and stepwise automated subset selection algorithms: frequency of obtaining authentic and noise variables. British Journal of Mathematical and Statistical Psychology. 45: 262-282.

Duarte Silva, A.P. Forthcoming. A "Leaps and Bounds" algorithm for variable selection in two-group discriminant analysis. IFCS-98 - Proceedings of Sixth Conference of the International Federation of Classification Societies. Rome, Italy.

Fisher, R.A. 1936. The use of multiple measurements in taxonomic problems. Annals of Eugenics. 7: 179-188.

Furnival, G.M. 1971. All possible regressions with less computation. Technometrics, 13: 403-408.

Furnival, G.M. and Wilson, R.W. 1974. Regressions by Leaps and Bounds. Technometrics, 16: 499-511.

Geisser, S. 1977. Discriminantion, allocatory, and separatory, linear aspects. In J. Van Ryzin (Ed.) Classification and Clustering. Academic Press, 301-330, New York, NY.

Hand, D.J. 1981. Discrimination and Classification. John Wiley. New York, NY.

Hocking, R.R. and Leslie R.N. 1967. Selection of the best subset in regression analysis. Technometrics. 9: 531-540.

Hottelling, H. 1936. Relations between two sets of variates. Biometrika. 28: 321-377.

Huberty, C.J. 1994. Applied Discriminant Analysis, New York, NY: Wiley.

Huberty, C.J. and Wisenbaker, J.M. 1992. Variable importance in multivariate group comparisons. Journal of Education Statistics, 17: 75-91.

Kobilinsky, A. 1990. Analyse factorielle discriminante. In Gilles Celeux (Ed). Analyse discriminant sur variables continues. INRIA, 65-80.

Kuk, A.Y.C. 1984. All subsets regression in a proportional hazard model. Biometrika. 71: 587-592.

Lachenbruch, P.A. 1968. On Expected Probabilities of Misclassification in Discriminant Analysis, Necessary Sample Size, and a Relation with the Multiple Correlation Coefficient. Biometrics, 24: 823-834.

Lachenbruch, P.A. and Mickey, M.R. 1968. Estimation of error rates in discriminant analysis. Technometrics. 10: 1-11.

Lawless, J. and Singhal, K. 1978. Efficient screening of nonnormal regression models. Biometrics. 34: 318-327.

Masson, J.P. 1990. Discrimination et analyse de variance. In Gilles Celeux (Ed). Analyse discriminant sur variables continues. INRIA, 81-99.

McCabe, G.P. 1975. Computations for Variable Selection in Discriminant Analysis. Technometrics, 17: 103-109.

McCulloch, R.E. 1986. Some remarks on allocatory and separatory linear discrimination. Journal of Statistical Planning Inference. 14: 323-330.
McHenry, C.E. 1978. Computation of a best subset in multivariate analysis. Applied Statistics. 27: 291-296.
McKay, R.J. and Campbell, N.A. 1982a. Variable Selection Techniques in Discriminant Analysis I. Description. British Journal of Mathematical and Statistical Psychology. 35: 1-29.
McKay, R.J. and Campbell, N.A. 1982b. Variable Selection Techniques in Discriminant Analysis II. Allocation. British Journal of Mathematical and Statistical Psychology. 35: 30-41.
McLachlan, G. J. 1974. An Asymptotic Unbiased Technique for Estimating the Error Rates in Discriminant Analysis. Biometrics. 30: 239-249.
McLachlan 1986. Assessing the performance of an allocation rule. Computers and Mathematics with Applications. 12A: 261-272.
McLachlan, G. J. 1992. Discriminant Analysis and Statistical Pattern Recognition, New York, NY: Wiley.
Miller, A.J. 1984. Selection of subsets of regression variables (with discussion). Journal of the Royal Statistical Society. *A.* 147: (3), 389-425.
Miller, A.J. 1990. Subset Selection in Regression. Chapman and Hall.
Murray, J.D. 1977. A cautionary note on selection of variables in discriminant analysis. Applied Statistics. 26: 246-250.
Rao, C.R. 1952. Advanced Statistical Methods in Biometric Research. John Wiley. New York, NY.
Rao, C.R. 1973. Linear Statistical Inference and its Applications, *2nd Ed.* John Wiley. New York, NY.
Rozeboom, W.W. 1965. Linear correlation between sets of variables. Psychometrika. 30: 57-71.
Schervish, M.J. 1981. Asymptotic expansions for correct classification rates in discriminant analysis. Annals of Statistics. 9: 1002-1009.
Seber, J.A.F. 1984. Multivariate Observations. John Wiley. New York, NY.
Shaffer, J.P. and Gillo, M.W. 1974. A multivariate extension of the correlation ratio. Educational and Psychological Measurement. 34: 521-524.
Snapinn, S.M. and Knoke, J.D. 1989. Estimation of error rates in discriminant analysis with selection of variables. Biometrics. 45: 289-299.
Turlot, J.C. 1990. Sélection des prédicteurs et estimation des taux d'erreur de classement en discrimination linéaire. In Gilles Celeux (Ed). Analyse discriminant sur variables continues. INRIA, 51-63.

NOTES:

(1) Generalized variance is understood as the determinant of a of variance-covariance matrix (Anderson 1958).
(2) In reality McCabe presents his approach in terms of trying to identify subsets resulting in low values for Λ. In this context, as McCabe rightly recognizes, Λ should not be interpreted as a test statistic, but simply as an index of "group proximity". In this article, in order to stress this point, it was chosen to describe McCabe approach in terms of η^2 which, contrary to Λ, is usually interpreted as an index.
(3) Actually, using the known relation $W = (N-k)\,V$, where V denotes Hotelling-Lawley statistic concercing the hypothesis of equal population means across groups, W could also be expressed as a sum of $k-1$ (instead of k) quadratic forms.